Thermodynamics of Protein Folding Studied by Umbrella Sampling along a Reaction Coordinate of Native Contacts

Hamed Meshkin and Fangqiang Zhu*

Department of Physics, Indiana University Purdue University Indianapolis, 402 N. Blackford St., Indianapolis, Indiana 46202

E-mail: fzhu0@iupui.edu

Abstract

Spontaneous transitions between the native and non-native protein conformations are normally rare events that hardly take place in typical unbiased molecular dynamics simulations. It was recently demonstrated that such transitions can be well described by a reaction coordinate, Q, that represents the collective fraction of the native contacts between the protein atoms. Here we attempt to use this reaction coordinate to enhance the conformational sampling. We perform umbrella sampling simulations with biasing potentials on Q for two model proteins, Trp-Cage and BBA, using the CHARMM force field. Hamiltonian replica exchange is implemented in these simulations to further facilitate the sampling. The simulations appear to have reached satisfactory convergence, resulting in unbiased free energies as a function of Q. In addition to the native structure, multiple folded conformations are identified in the reconstructed equilibrium ensemble. Some conformations without any native contacts nonetheless have rather compact geometries and are stabilized by hydrogen bonds not present in the native

This is the author's manuscript of the article published in final edited form as:
Meshkin, H., & Zhu, F. (2017). Thermodynamics of Protein Folding Studied by Umbrella Sampling along a Reaction Coordinate of Native Contacts. Journal of Chemical Theory and Computation, 13(5), 2086–2097. https://doi.org/10.1021/acs.jctc.6b01171

structure. Whereas the enhanced sampling along Q reasonably reproduces the equilibrium conformational space, we also find that the folding of an α-helix in Trp-Cage is a slow degree of freedom orthogonal to Q and therefore cannot be accelerated by biasing the reaction coordinate. Overall, we conclude that whereas Q is an excellent parameter to analyze the simulations, it is not necessarily a perfect reaction coordinate for enhanced sampling, and better incorporation of other slow degrees of freedom may further improve this reaction coordinate.

1 Introduction

The function of a protein is determined by its three-dimensional structures.[1,2] Many proteins adopt a specific folded conformation, referred to as the native structure, under physiological conditions. Thermodynamically, the native structure typically corresponds to a minimum in the free energy surface. Early theoretical analysis suggested that the native structure would obey the minimal frustration principle,[3,4] and recent simulation studies further revealed that the native structure also serves as a kinetic hub that connects multiple highly distinct non-native conformations.[5] Indeed, the native structure is not necessarily the only conformation adopted by a protein, and there may exist an equilibrium between the native structure and the non-native (such as disordered and extended) conformations. The thermodynamics and kinetics for the transitions between the native and the non-native protein structures, such as the folding rate,[6] the transition state,[7] and the intermediates states,[8] have been extensively studied for decades.

Computational methods such as molecular dynamics (MD) simulations[9] are powerful tools to complement protein folding experiments. Among all the MD methods, the most straightforward approach is to directly simulate a protein in its natural environment and observe the spontaneous transitions between the native and the non-native conformations. If the simulation is long enough such that a statistically sufficient number of transitions occur, all thermodynamic and kinetic quantities of protein folding can be directly obtained from

the simulation trajectory. Thanks to the breakthrough in specialized computer hardware and algorithm, all-atom simulations of millisecond time scale have been achieved,[10,11] which allowed direct observation of folding/unfolding transitions for a number of small proteins with relatively fast kinetics. Alternatively, a variety of enhanced sampling methods have been applied to simulate protein folding.[9,12,13] Some of these methods, such as umbrella sampling (US)[14,15] and metadynamics,[16] employ non-Boltzmann sampling with biasing potentials to accelerate the transitions over the energy barriers. Similar acceleration can also be achieved, e.g., in weighted ensemble simulations,[17,18] by generating multiple replicas to enhance the sampling in regions with low equilibrium probabilities. In all of the methods above, the unbiased equilibrium thermodynamics can be reconstructed from the simulation trajectories, based on rigorous theories in statistical mechanics. In addition, serial or parallel tempering[19] can be employed in methods such as replica exchange MD (REMD) simulations,[20,21] in which multiple replicas are run in parallel and periodically attempt to exchange their temperatures or biasing potentials.[22–25]

An exact protein conformation must be described in a multidimensional space. Indeed, the conformational space for proteins has been successfully described by Markov state models.[26] Alternatively, in many cases it is also desirable to project the high-dimensional protein conformations onto a single reaction coordinate (or order parameter) to simplify the analysis. Once such a reaction coordinate is defined, its equilibrium probability distribution can be determined from the equilibrium ensemble of the protein conformations and will correspond to a free energy as a function of the reaction coordinate. With a "good" reaction coordinate for protein folding, the associated free energy would not only clearly distinguish the native and the non-native states, but also reflect the kinetic barrier for the transitions.

Many common reaction coordinates for protein folding are based on the fraction of native contacts.[27] A contact is usually defined as a pair of residues that are spatially close (shorter than some cut-off distance) but not in sequence proximity, and all such contacts in the native structure constitute the set of native contacts. One can then examine how many of the native

contacts are present or absent in any given conformation based on the inter-residue distances. As a simple criterion, a Heaviside step function[28] can be used to map a distance to a contact number, which can be either 0 or 1 as determined by the cut-off distance. Other criteria assign a non-integer contact number between 0 and 1 using a continuous function of the distance, such as Gaussian[29–31] or Fermi-Dirac distribution functions.[11,15,32–35] The sum of the contact numbers in the given conformation, as a fraction of the maximum possible total number (as in the native structure), can then be defined as the reaction coordinate, with a value close to 1 and 0 representing the native and the non-native states, respectively. Alternative to the native contacts, reaction coordinates can also be defined based on dihedral angles,[36–38] native hydrogen bonds,[15] the number of core water molecules,[39,40] as well as holistic parameters such as radius of gyration[28] and root-mean-square deviation (RMSD).[41]

Recently, Best et al.[42] analyzed the trajectories of millisecond-long unbiased MD simulations[11] of some small proteins and concluded that a reaction coordinate based on the collective fraction of native contacts characterizes the folding/unfolding transitions remarkably well.[42] In principle, once a good reaction coordinate is identified, enhanced sampling along that coordinate could provide the conformational thermodynamics in a potentially more efficient way compared to the straightforward unbiased simulations. Here we test this strategy by performing US along the reaction coordinate mentioned above, as similarly done in some earlier studies.[15,43–45] Our all-atom simulations are performed with explicit solvent, and we employ the Hamiltonian REMD technique[22] to facilitate the US[14,15] in this study. We use two small proteins, Trp-Cage[46] and zinc finger motif (BBA),[47] as the test cases here. Trp-Cage is a 20-residue protein that can fold rapidly to a stable structure. BBA is a 28-residue protein with a native structure that consists of two β-sheets and one α-helix. Both proteins have been extensively studied in previous simulations.[11,21,46–49] We determine the free energy profile and reconstruct the equilibrium ensemble for each protein from the simulations here.

Our simulations serve as a case study for using the reaction coordinate based on the native

contacts to sample protein conformations. Through detailed analysis, we demonstrate the effectiveness and the problems with this approach. Although we specifically adopted US in this study, we note that many other enhanced sampling methods also require a pre-determined reaction coordinate and would have similar problems with the folding reaction coordinate examined here.

2 Methods

In this study, we focus on the folding of two proteins, Trp-Cage[46] and BBA,[47] which have also been extensively studied in previous simulations.[5,11,21,41,46–49] In particular, Lindorff-Larsen et al.[11] performed long unbiased simulations on the two proteins, and Best et al.[42] analyzed the simulation trajectories using a reaction coordinate representing the collective fraction of native contacts. Here we take the reaction coordinate above and perform US[14,15] simulations with Hamiltonian Replica Exchange Molecular Dynamics (HREMD)[22] to reproduce the equilibrium ensemble for the proteins. The computational details are provided below.

1. System Setup. Both of our simulation systems are similar to the ones used in Lindorff-Larsen et al.[11] The first protein is a Trp-Cage mutant, denoted as TC10b (PDB: 2JOF[50]), with the sequence DAYAQWLADGGPSSGRPPPS. In comparison to the wild type, residue 8 in the sequence is mutated from LYS to ALA.[11] The simulation system consists of the protein in a solution of 1639 water molecules and 65 mM NaCl. The total number of atoms in the Trp-Cage simulation system is 5230. The second protein, BBA (PDB: 1FME[47]), with the sequence EQYTAKYKGRTFRNEKELRDFIEKFKGR, was solvated with 2978 water molecules and four Chloride ions. The simulation system for BBA consists of a total of 9442 atoms. We adopted the standard protonation state at pH 7 for all residues of the two proteins. For both proteins, the first frame in the PDB file was taken as the native structure in this study.

We adopted the CHARMM (Ver. c36, released in December 2013) protein force field[32,51,52]

and the TIP3P water model[53] in this study. The MD simulations were carried out using the NAMD2 program[54] with a time step of 2 fs and in the NPT ensemble with the periodic boundary conditions. A constant pressure of 1 atm was obtained by applying the Nose-Hoover Langevin piston method,[55] and a Langevin thermostat with a damping coefficient of 1 ps^{-1} was used to maintain the constant temperature of the system. The SHAKE[56] and SETTLE[57] algorithms were used to maintain rigid bonds involving all hydrogen atoms. We used a 12 Å cut-off for non-bonded interactions, with a smooth switching function starting at 10 Å. Full electrostatics was calculated every 4 fs using the particle mesh Ewald (PME) method.[58]

The two systems were first minimized and equilibrated for a total of 10 ns. In the equilibration phase, the temperatures of the Trp-Cage and the BBA systems were 290 K and 325 K, respectively, although Trp-Cage was simulated at two additional temperatures as well, as will be described later.

2. Reaction Coordinate. We adopt the same reaction coordinate in Best et al.[42] based on the fraction of native contacts. The set of native contacts is defined from the native structure. Specifically, a pair of heavy atoms (i, j) in residues R_i and R_j is counted as a native contact if $|R_i - R_j| > 3$ and the interatomic distance r_{ij}^0 in the native structure is smaller than 4.5 Å. In our case, the number of native contacts identified from the crystal structure is $N = 156$ and $N = 279$ for Trp-Cage and BBA, respectively. Assuming that the atom pair (i, j) is one of the native contacts, we use $r_{ij}(X)$ to denote the distance between the two atoms in a given protein conformation X. The reaction coordinate Q for any conformation X is then determined by the distances for the N pairs of atoms in this conformation:[42]

$$Q(X) = \frac{1}{N} \sum_{ij}^{N} \frac{1}{1 + \exp[\beta(r_{ij}(X) - \lambda r_{ij}^0)]}, \tag{1}$$

with $\lambda = 1.8$ and a smoothing parameter $\beta = 5.0$ $Å^{-1}$. The summand in the equation above is effectively a pairwise contact strength that approaches 1 when the distance r_{ij} is small

and approaches 0 when r_{ij} is large, thus quantifying the degree of contact between the two atoms. The reaction coordinate (Q) is the average over all pairwise contact strengths, thus representing the collective fraction of the native contacts present in a given conformation. A value of Q close to 1 indicates that the protein is in the native state because all of the native contacts are intact. In contrast, $Q \sim 0$ corresponds to completely non-native structures with all the native contacts broken.

3. Umbrella Sampling Simulations. We employed a total of 32 umbrella windows. The biasing potential in window i is in the harmonic form:

$$U_i(X) = \frac{K}{2}[Q(X) - q_i]^2, \tag{2}$$

in which $i = 1, \ldots, 32$. The spring constant K was taken to be 1400 kcal/mol for all the simulations in this study, and q_i is the center of each harmonic biasing potential. The values of q_i ($i = 1, \ldots, 32$) cover the range from 0 to 1 with a uniform spacing of 1/31.

To start the US simulations, we need a set of initial conformations with the reaction coordinate close to the q_i in each window. One common method to generate a diverse set of conformations is to run an equilibrium simulation at high temperatures.[15] Here we instead adopted pulling simulations, similar to the steered molecular dynamics,[59] for this purpose. Specifically, we performed a simulation to drive the system from the native state ($Q \sim 1$) to the non-native state ($Q \sim 0$), by sequentially applying the 32 umbrella potentials for 0.4 ns each. The simulation thus lasted for a total of 12.8 ns. From this simulation trajectory, frames with the reaction coordinate close to each q_n were then selected as the initial coordinates for the respective umbrella window.

In the US, the umbrella windows were sampled by the same number of individual simulations (each referred to as a replica), and HREMD[22] was implemented to allow two neighboring windows to swap their replicas. The exchange was attempted every 200 time steps (i.e., 0.4 ps). Suppose that umbrella windows i and j are a pair of neighbors, and that at the

time of an exchange attempt, the current reaction coordinates are Q_i and Q_j, respectively. A swap would thus change the combined Hamiltonian by $\Delta E = K(Q_i - Q_j)(q_i - q_j)$, in which q_i, q_j, and K are from the harmonic biasing potential (Eq. 2). We accept the exchange with a probability of $\min[\exp(-\frac{\Delta E}{k_B T}), 1]$ according to the Metropolis Criterion.[22] If the exchange is accepted, the two umbrella windows will swap their replicas, thus effectively exchanging the system microstates (coordinates, velocities, etc.).

We performed a total of four sets of US simulations, including the Trp-Cage system at 270 K, 280 K, and 290 K, and the BBA system at 325 K. Each simulation of Trp-Cage was run for 3.00 μs per window, or a total of 96.00 μs for the 32 windows. The simulation of BBA was run for 1.01 μs per window, or 32.32 μs in total. The initial coordinates for the Trp-Cage simulation at 290 K and the BBA simulation were taken from the pulling simulations described earlier. The last frames of the Trp-Cage simulation (290 K) were then used to initiate the US simulations at 280 K and 270 K.

4. Analysis. The second half of the trajectories was used for the analysis of each simulation. Due to replica exchange, each umbrella window may be sampled by different replicas at different times of the simulation. We thus first reassembled the trajectories for each umbrella window. From these trajectories, we constructed the histograms of Q for each window, using a uniform bin width of $\Delta Q = 1.1 \times 10^{-4}$ for the Trp-Cage simulations and $\Delta Q = 2.0 \times 10^{-4}$ for the BBA simulation. Then the weighted histogram analysis method (WHAM)[60,61] was used to calculate the equilibrium free energy as a function of Q.

With the equilibrium probability distribution of Q and the trajectories from the US simulations, we can reconstruct the equilibrium ensemble and obtain the probability distribution for any given parameter R, such as RMSD or radius of gyration. Specifically, we first group all frames in the simulation trajectories according to their values of Q. For each set of frames with the same Q, we construct the histogram for R as an estimate for the conditional probability $P(R|Q)$. In addition, $P_Q(Q)$, the marginal distribution for Q, is directly obtained from WHAM or the free energy $G(Q)$. The joint equilibrium probability for R and Q is therefore

given by $P(R,Q) = P(R|Q)P_Q(Q)$.

3 Results

As described in Methods, we performed US simulations with HREMD[22] on the Trp-Cage[46] and BBA[47] systems, using a reaction coordinate[42] Q based on the native contacts. The Trp-Cage system was simulated at three different temperatures.

3.1 Equilibrium distributions along the reaction coordinate

Figure 1a shows the free energy profiles as a function of the reaction coordinate Q, obtained from the US simulation trajectories by WHAM.[60,61] The statistical errors were estimated from the uncertainties of the mean forces at each window.[60] Overall, the free energy profiles here do not appear to describe a typical two-state system that has two major metastable states separated by a prominent energetic barrier. Instead, the profiles feature multiple minima and peaks with magnitudes not significantly larger than k_BT, thus indicating a continuous spectrum of intermediate conformations at equilibrium. In general, the locations of the major free energy barriers in our profiles are qualitatively similar to those reported by Best et al.[42] for the long unbiased simulations,[11] although the magnitudes are not in good agreement. We caution that due to the different force fields adopted, the two studies are not expected to yield similar quantitative results. Figure 2a shows the cumulative distribution function (CDF) that integrates the equilibrium probability along Q. For Trp-Cage at the three temperatures, the free energies and the CDFs show that the equilibrium populations of the native (with large Q) and the non-native (with small Q) states are roughly comparable. For BBA at 325 K, in contrast, the vast majority of the equilibrium population is in the non-native state.

Any MD sampling has to start with some initial coordinates of the system, and convergence is only achieved when the "memory" has been completely lost and the results become

independent of the initial state. In our case here, although we discarded the first half of the trajectories in our analysis, slow equilibration in degrees of freedom orthogonal to the reaction coordinate could still potentially give rise to convergence issues. For umbrella sampling, one way to detect such issues is to examine the consistency between the histograms from neighboring windows. As described in Ref. 60, the two neighboring histograms should ideally predict a consensus probability distribution for the overlapping region. Insufficient sampling of the orthogonal degrees of freedom, or hysteresis, often manifests itself as an inconsistency between the histograms.[60] Therefore, for every pair of adjacent umbrella windows, we compared their consensus probability distributions (under a common potential) reconstructed from the two histograms. For such comparison, we adopted the inconsistency coefficient $\theta_{(i,i+1)}$ defined in Ref. 60 based on the Kolmogorov-Smirnov test. A θ value much larger than 1 would indicate an abnormal inconsistency between the two histograms. Figure 2b shows that all θ values from our simulations are below 1.05, and therefore no major inconsistency is detected. This analysis thus suggests that the calculated statistical errors here are reasonable estimates for the actual sampling errors.

HREMD[22] was implemented in our simulations, with the exchange rates between neighboring windows in the range of 20%-40%. In this scheme, the biasing potential on each replica undergoes a discrete random walk during the simulation.[62] The behavior of such random walk, quantified by parameters such as the transmission factors,[62] could also potentially reveal regions with slow relaxation in the degrees of freedom orthogonal to the reaction coordinate.[62] The calculated transmission factors for our simulations did not exhibit significant variations[62] across different regions of Q, and thus did not indicate any particularly problematic region for the sampling. Figure 3 shows the umbrella windows sampled by each replica during the simulations. The sampled ranges for the individual replicas are clearly very different. The majority of the replicas visited a substantial range of the umbrella windows, with few covering almost the entire Q-range while some only covering a narrow section. It is well known that due to the effect of replica sorting,[62] the replicas in HREMD simulations

tend to be trapped in local regions.

The ultimate validation of an enhanced sampling method (such as US) would be a direct comparison to ideally long unbiased simulations. Although millisecond simulations[11] were not affordable here, we performed unbiased simulations from the native state of Trp-Cage at 280 K as an additional test. Specifically, we took a total of 32 frames in the US trajectories, with the reaction coordinate Q ranging from 0.94 to 0.98. From each frame, we initiated an unbiased simulation (without any restraint) for 344 ns. The histograms of each simulation from the second half (172 ns) of the trajectory are shown in figure 4 (*dotted* lines). Remarkably, the histograms from these individual simulations are still significantly different from each other after 344 ns, thus indicating that the equilibration is not very fast even when the protein is near the local free energy minimum for the native conformation, presumably due to the effects of other degrees of freedom. Whereas the protein in most unbiased simulations stayed in the native conformation during the 344 ns, we also observed a single spontaneous partial unfolding transition in one simulation, with the protein converted to some intermediate conformations with $Q \sim 0.4$. Overall, despite the large variations among the individual histograms, their average is in reasonable agreement with the prediction from the US simulations (Figure 4).

In principle, with the knowledge of the free energy and the diffusion coefficients along the reaction coordinate, one may further obtain the kinetics of the transition.[63,64] Although we performed some additional US simulations to calculate the diffusion coefficients,[65] the statistical uncertainties appeared to be very large. Furthermore, the thermodynamics here does not indicate a two-state transition, as mentioned earlier. Therefore, we did not further estimate the folding/unfolding rates for the transition as in other studies.[63,64]

3.2 Energetics of the conformational space

The Gibbs free energy (G) can be decomposed as the enthalpy (H) and the entropy (S): $G = H - TS$. Our US simulations could provide these thermodynamic quantities for different

conformational states (described by Q). As discussed earlier, the free energy as a function of Q was calculated by WHAM.[60,61] Furthermore, we calculated the enthalpy for each frame in the simulation trajectories as $H = U + PV$, in which U is the potential energy for the underlying atomic interactions, V is the volume of the simulation system, and P is the pressure. Under the constant pressure of 1 atm, the variations in the PV term are much smaller than in the potential energy U. We took the average for all snapshots with the same Q as the enthalpy value at that Q. The entropy was then determined from the difference between the free energy and the enthalpy.

The enthalpy and entropy of each system are shown in figure 1, b and c, along with the free energy. In general, the variations in the enthalpy here are larger than in the free energy. For BBA, as expected, the minimum enthalpy is at large Q representing the native state. For Trp-Cage, surprisingly, the enthalpy for the native state is actually not the global minimum. Instead, the enthalpy minimum for Trp-Cage is at $Q \sim 0.5$, thus suggesting that some intermediate conformations, as will be described in more details later, actually have even more favorable potential energies than the native structure.

We also attempted to calculate the heat capacity for the conformations at different Q, obtained from the equilibrium energy fluctuation. However, the statistical uncertainties in this calculation are too large to reveal any clear difference of the heat capacity across the range of Q.

3.3 Stability of the native contacts

The Trp-Cage crystal structure consists of a short α-helix (residues 2–9) and a Polyproline-II segment (residues 16–19) connected by a loop (residues 10–15) that contains a 3_{10}-helix. The indole ring of the tryptophan residue (W6) is located at the center of the protein and makes contact with all of the three segments. Our simulation trajectories reveal different degrees of stability for the three segments, as shown in figure 5 for the average fraction of the native contacts between each pair of protein residues for conformations at different Q. The contact

maps for all three temperatures are quite similar, with the ones for 270 K and 290 K shown in the figure. Whereas the reaction coordinate Q is essentially an aggregate of the pairwise contacts, the maps indicate that the individual contact strengths do not simply increase linearly with Q from the non-native to the native states. Instead, the pairwise contacts are formed in different stages, thus implying different stabilities for the three segments. In particular, the α-helix appears to have the most stable secondary structure. At a relatively low Q (0.3 or 0.4), the signature contacts within the α-helix already become prominent. In contrast, contacts involving the Polyproline-II and the loop segments appear to be less stable. For example, the native contacts between W6 and those two segments only start to form at $Q = 0.7$. Finally, some native contacts are quite weak even in the highly native conformations. For instance, the average contact strength for the D9-R16 salt bridge is smaller than 0.3 among the conformations at $Q = 0.9$.

Some insight on the relative stability can also be gained from the spontaneous transition away from the native structure observed in the unbiased simulation described earlier. In this transition, the α-helix remained essentially unchanged whereas the loop and the Polyproline-II segment underwent large deviations from the initial native conformation. In the end of the partial transition, the protein is in a partly native conformation with an intact α-helix. This observation is consistent with our conclusion of a more stable α-helix and suggests that the unfolding of the α-helix would be the last step in reaching the completely non-native conformation.

3.4 Radius of gyration

The free energies discussed above are directly related to the marginal probability distribution of Q at equilibrium, with all other degrees of freedom integrated out. It is thus possible that highly distinct conformations are mapped to a same value of Q. In the meantime, other parameters can be introduced to represent the equilibrium ensemble from different angles. As described in Methods, we can project the equilibrium ensemble onto any parameters and

obtain the joint probability distribution. The free energy as a function of those relevant parameters may then reveal conformational states that otherwise cannot be distinguished by Q alone.

One relevant order parameter is the radius of gyration, Rg, which measures the geometric extendedness of the protein conformation.[4,28] Figure 6 shows two-dimensional free energies as a function of Q and Rg, obtained from their joint probability distribution in the equilibrium ensemble. Qualitatively, the free energy maps for all simulations exhibit some common features. At large Q, the protein is in the native state, and Rg is therefore narrowly distributed around the value for the crystal structure. As Q decreases, the sampled range of Rg becomes increasingly larger, indicating the presence of more extended conformations. However, all major free energy minima, regardless of Q, are located at small values of Rg, and therefore the vast majority of the equilibrium population has Rg values similar to the crystal structure. Even for the non-native state near $Q = 0$ with all the native contacts completely lost, highly extended conformations (with large Rg) only represent a very small fraction of the population. These observations indicate that the non-native states here, albeit completely different from the crystal structure, are still folded in fairly compact geometries.

The two-dimensional free energy maps reveal a number of metastable conformations that are not clearly distinguishable in the one-dimensional profile. Some of the conformations are shown in figure 6 for Trp-Cage at 270 K. At $Q \sim 1$, conformation A is the native state as defined by the crystal structure. Around $Q \sim 0.5$, conformations B–D are partly native conformations with the α-helix similar to the crystal structure but the loop region highly different, especially for conformations C and D. In conformation B, the R16 guanidinium group simultaneously forms salt bridges with the carboxylate groups of both D1 and D9. In conformation C, the Polyproline-II segment contacts the α-helix, and the W6 indole ring forms an H-bond with the backbone carbonyl group of P12 or S13. Conformation D is similar to conformation C, except that the W6 indole ring H-bonds with the backbone carbonyl group of S14, G15 or R16, or with the sidechain of S13. At low values of Q, conformations

E–I correspond to completely non-native structures. Among them, conformation I is a fully extended structure with the maximum Rg (17 Å). The equilibrium population of this extended conformation, however, is small in comparison to other non-native conformations. Those conformations (E–H) have lost almost all of the native contacts but nonetheless are nearly as compact (with Rg 7–9 Å) as the native structure (with Rg 6.9 Å). They are mainly stabilized by a different set of H-bonds that are not present in the native structure, as will be further discussed later.

Overall, the Q-Rg free energy maps (Figure 6) of Trp-Cage at the three temperatures are qualitatively similar. The average Rg in the entire equilibrium ensemble is 8.1 Å, 7.8 Å, and 8.0 Å at 270 K, 280 K, and 290 K, respectively. However, the free energy minima corresponding to the distinct conformations discussed above are most prominent at 270 K, although those conformations can indeed be found (with somewhat lower probabilities) in the equilibrium ensembles at 280 K and 290 K as well. In addition, the relative free energy at small Q for 270 K is lower than that for the other two temperatures, thus indicating that the equilibrium population of the non-native conformations (such as the fully extended conformation) is higher at 270 K. For protein BBA, the two-dimensional free energy map indicates that the non-native state (with low Q) is more predominant than the other states (Figure 6), also consistent with its one-dimensional $G(Q)$ profile (Figure 1a). Similar to the case of Trp-Cage, the majority of the non-native BBA conformations are relatively compact, with Rg comparable to its native structure.

3.5 Hydrogen bonds

H-bonds are believed to play important roles in the stability of protein conformations.[39,66] We identified all H-bonds in the simulation trajectories, using a criterion that the donor-acceptor (which can be N or O atoms) distance be smaller than 4.0 Å and the donor-H-acceptor angle be larger than 140°. The identified H-bonds are classified as native hydrogen bond (NHB) or non-native hydrogen bond (N-NHB), depending on whether they are present in the native

crystal structure or not. Using the criteria above, there are a total of 12 NHBs in the crystal structure. One NHB is actually a salt bridge between the guanidinium group of R16 and the carboxylate group of D9. Another NHB is between the sidechain indole ring of W6 and the backbone carbonyl group of R16. The other 10 NHBs are between the backbone amide N−H and the carbonyl C=O groups in residues 1–15.

Figure 7 shows two-dimensional free energy maps determined from the joint probability distribution of Q and the number of NHBs or N-NHBs in the equilibrium ensemble of Trp-Cage. As expected, the number of NHBs strongly correlates with Q, the fraction of the native contacts. For the free energy basin corresponding to the native state, most conformations have at least 7 NHBs. There are typically 4–6 NHBs in the intermediate conformations with Q between 0.3 and 0.7, whereas the non-native conformations have no more than 3 NHBs. In contrast, the number of N-NHBs does not appear to depend on Q. Even for the completely non-native conformations with $Q \sim 0$, the number of N-NHBs is similar to that in the native conformations. As discussed earlier, most conformations at $Q \sim 0$ still have folded geometries that are almost as compact as the native structure. Results here thus suggest that these compact non-native conformations are stabilized by different sets of H-bonds that are not present in the native structure.

3.6 Folding of the α-helix in Trp-Cage

As described earlier, the α-helix at the N-terminal of Trp-Cage is largely intact in the partly native conformations, thus suggesting that the forming of this α-helix would be an important step in the folding transition. We calculated the RMSD values (denoted as RMSD_{hx}) of the C_α atoms in the α-helix for all conformations in the simulation trajectories, using the native α-helix structure as the reference. Figure 8 displays the two-dimensional free energy maps as a function of Q and RMSD_{hx} for the equilibrium ensemble. Interestingly, the free energies exhibit a more prominent two-state signature along the RMSD_{hx} parameter than along Q. There are two major minima along RMSD_{hx}: the minimum at $\mathrm{RMSD}_{hx} \sim 0$ corresponds to

the folded α-helix (such as conformations A–D in figure 6), and the minimum around 3–5 Å corresponds to the completely unfolded helix (such as conformations E–I in figure 6). Some intermediate conformations (with RMSD_{hx} around 2 Å) of a partially folded α-helix also exist, but only with minor populations. Overall, there thus exists an energetic barrier along RMSD_{hx}, as also identified in an earlier REMD study.[67] The free energy maps also show that the transitions along RMSD_{hx} would occur when Q is around 0.3. Some folding/unfolding transitions of the α-helix are described in Supporting Information.

4 Discussion

In this study, using a reaction coordinate representing the collective fraction of the native contacts, we carried out US[14] simulations in combination with HREMD[22] to sample the protein conformational space. Overall, the free energy calculation (Figure 1a) appears to have converged, and the consistency test (Figure 2b) suggests that the statistical errors in the free energy have been reasonably estimated. The equilibrium ensemble of protein conformations thus appears to be satisfactorily reconstructed from these simulations.

The reconstructed equilibrium ensembles reveal multiple folded conformations for the two proteins here, Trp-Cage and BBA. The reaction coordinate Q only quantifies the resemblance to the native structure but does not describe the compactness of the conformation. In fact, the non-native state does not merely consist of disordered or extended conformations. Even at $\mathrm{Q} \sim 0$, with all native contacts completely broken, the majority of the populations are still comprised of well-defined conformations almost as compact as the native structure (Figure 6), and these folded conformations are stabilized by some H-bonds (Figure 7) not present in the native structure. For Trp-Cage, some alternatively folded conformations have even lower enthalpy than the native structure (Figure 1b). In the presence of such conformations,[68] therefore, the conformational space would not be described by a simple two-state model with a folded conformation and an unfolded state of disordered conformations.

For Trp-Cage, the α-helix at the N-terminal plays an important role in the folding of this protein. UV resonance Raman spectroscopy[69] detected in the unfolded ensemble the presence of compact intermediate conformations with the intact α-helix, and concluded that the Trp-Cage is not a two-state folder.[69] Infrared spectroscopy also indicated that the α-helix is fully formed in the folding transition state.[70] These conclusions were further supported by recent simulations.[71] Our simulations here showed that the α-helix is more stable than other parts of the protein (Figure 5) and is largely intact in the intermediate conformations at $Q \sim 0.5$. Furthermore, the spontaneous partial unfolding transition in one of our unbiased simulations showed that the α-helix remained intact when other parts of the protein deviated from the native conformation. Therefore, our simulations are fully consistent with the previous findings that the α-helix is formed at the early folding stage, although we caution that the Trp-Cage sequences in those experimental studies[69,70] are slightly different from ours. Importantly, our reconstructed equilibrium ensemble revealed that the transition between the folded and unfolded α-helix is almost orthogonal to the reaction coordinate Q (Figure 8). Consequently, the restraint on Q in the US simulations cannot enhance the sampling of the α-helix conformations, which would thus compromise the sampling efficiency and contribute to the statistical errors in the free energy. Furthermore, the one-dimensional free energy as a function of Q does not reflect the energetic barrier between the folded and unfolded conformations of the α-helix (Figure 8). In fact, the folding of the α-helix resembles a two-state process more than the folding of the entire Trp-Cage does, as also noted in previous experiments.[69]

Trp-Cage at various temperatures has been studied in NMR experiments.[68,72] Here we carried out simulations at three different temperatures (270 K, 280 K, and 290 K) for this protein. Whereas the reconstructed equilibrium ensembles at these temperatures are qualitatively similar to each other, it is notable that the non-native state turns out to have a higher equilibrium probability at the lowest temperature (270 K) than at the other temperatures (Figure 2a). This somewhat unexpected result may be attributed to several factors. First,

given the relatively small magnitude of the free energies here, the statistical errors in our calculation are relatively large. Consequently, the differences in the calculated equilibrium probabilities at the three temperatures are not much larger than the estimated statistical uncertainty. More importantly, as discussed earlier, the equilibrium ensembles consist of multiple folded conformations. Some alternatively folded conformations are enthalpically even more favorable than the native structure (Figure 1b). Consequently, lowering the temperature is not guaranteed to shift the equilibrium toward the native structure and away from other folded conformations. In fact, at the lowest temperature (270 K) here, the completely non-native conformations (at $Q \sim 0$) have even lower relative enthalpies, which could be responsible for their higher equilibrium populations than at the other temperatures. Finally, some Trp-Cage mutant was found to exhibit cold denaturation at low temperatures,[73,74] and this mechanism remains a possibility in our case as well.

Despite some qualitative agreement, our results considerably deviate from previous simulations.[11,42] Most notably, here some compact non-native Trp-Cage conformations have even lower enthalpies than the native structure does, which is clearly unexpected. For BBA, moreover, our free energy profile (Figure 1a) indicates that the non-native state is significantly more stable (by ∼4 kcal/mol) than the native state, which is also different from previous simulations.[11] Such discrepancies are most likely due to the force field issues. First, our version of the CHARMM36 force field was retrieved before the most recent updates for improving the sampling of disordered protein states. More importantly, oversampling of compact conformations has been identified as a common deficiency of some force fields,[75–78] and the high populations of compact conformations in our equilibrium ensemble may well be due to such artifacts. In addition, the CHARMM force field is known to over-stabilize the interaction between the guanidinium and the carboxylate groups,[79,80] thus very likely responsible for the unexpectedly low enthalpy for the Trp-Cage conformations at $Q \sim 0.5$, some of which (Figure 6, conformation B) are indeed stabilized by salt bridges between the ARG and ASP residues. Although the optimized CHARMM22* force field[79] appears to pro-

duce excellent results in folding simulations,[11] the predicted enthalpy for Trp-Cage still has a large discrepancy[11] with experiments. In light of such problems, it should be worthwhile to use the many available NMR data[68,72] on small model proteins such as Trp-Cage to validate and calibrate the force fields.[81]

As mentioned before, with a good reaction coordinate, many enhanced sampling methods, including the US simulations adopted in this study, can be applied to sample the protein conformations. Here we demonstrated that using Q as the reaction coordinate, US in combination with HREMD[22] could reasonably sample the protein conformational space and reconstruct the equilibrium ensemble. The efficiency of such methods relative to the unbiased simulations, however, clearly depends on the underlying kinetics. For the Trp-Cage with relatively fast transition rates here, given the aggregated simulation times one could alternatively obtain multiple spontaneous transitions in unbiased simulations. The advantage of the US approach is therefore not prominent here (other than a technical gain of much better parallel efficiency). However, the required sampling time for unbiased simulations may increase by many orders of magnitude for proteins with slow kinetics. Even for BBA, a fast-folding protein, because the system is not at the melting temperature[10,11] here, in unbiased simulations the protein would predominantly stay in the non-native state and the spontaneous transitions will be significantly less frequent, thus requiring much longer simulation times. In contrast, with a good reaction coordinate, the computational cost for the US[14,15] and other enhanced sampling methods would not be nearly as sensitive to the height and skewness of the underlying free energy, and they have been routinely used to calculate free energies with high barriers in many applications. Furthermore, unlike the temperature replica exchange simulations which typically require more replicas for systems of higher atom count (such as in the explicit-solvent simulations), the enhanced sampling methods based on a reaction coordinate can be readily applied to systems of any size.

On the other hand, the success of the US as well as many other methods critically depend on the quality of the adopted reaction coordinate. An ideal reaction coordinate

should ensure that all orthogonal motions can be well equilibrated within the simulation time. A poor reaction coordinate could severely compromise the sampling efficiency as well as cause other problems. The fraction of the native contacts, Q, appears to be a reasonable reaction coordinate, as we could generate the non-native states and reproduce the equilibrium distribution by applying restraints on Q alone in the simulations. On the other hand, Q is probably not always a perfect reaction coordinate for enhanced sampling, as we also identified slow equilibration of an orthogonal degree of freedom, i.e., the folding/unfolding of the α-helix, for the protein Trp-Cage here. In such cases, Hamiltonian replica exchange could somewhat alleviate the problem of slow orthogonal relaxations and facilitate the sampling along an imperfect reaction coordinate.[62] We also note that the identified problems with Q may be partly due to the force field issues discussed earlier, as Q was shown to be a very good reaction coordinate[42] to analyze folding simulations[11] using the CHARMM22* force field. Nonetheless, our finding in this study suggests that the reaction coordinate Q could be improved, e.g., by better incorporating the slow degrees of freedom representing the α-helix conformation for Trp-Cage, and that an improved reaction coordinate should further enhance the sampling efficiency.

Acknowledgement

All the simulations were performed on the Big Red II supercomputer at Indiana University, and all the analysis was carried out on a Linux cluster (Pyrite) at School of Science, IUPUI.

References

(1) Chan, H. S.; Zhang, Z.; Wallin, S.; Liu, Z. Cooperativity, local-nonlocal coupling, and nonnative interactions: principles of protein folding from coarse-grained models. *Annu. Rev. Phys. Chem* **2011**, *62*, 301–326.

(2) Dobson, C. M. Protein folding and misfolding. *Nature* **2003**, *426*, 884–890.

(3) Bryngelson, J. D.; Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **1987**, *84*, 7524–7528.

(4) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Struct., Funct., Bioinf.* **1995**, *21*, 167–195.

(5) Dickson, A.; Brooks III, C. L. Native states of fast-folding proteins are kinetic traps. *J. Am. Chem. Soc.* **2013**, *135*, 4729–4734.

(6) Plaxco, K. W.; Simons, K. T.; Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **1998**, *277*, 985–994.

(7) Alm, E.; Baker, D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 11305–11310.

(8) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and âĂIJen-routeâĂİ intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **2000**, *298*, 937–953.

(9) Klenin, K.; Strodel, B.; Wales, D. J.; Wenzel, W. Modelling proteins: Conformational sampling and reconstruction of folding kinetics. *BBA, Biochim.Biophys. Acta, Proteins Proteomics* **2011**, *1814*, 977–1000.

(10) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341–346.

(11) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517–520.

(12) Adcock, S. A.; McCammon, J. A. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev. (Washington, DC, U. S.)* **2006**, *106*, 1589–1615.

(13) Abrams, C.; Bussi, G. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy* **2013**, *16*, 163–199.

(14) Kästner, J. Umbrella sampling. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 932–942.

(15) Shea, J.-E.; Brooks III, C. L. From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem* **2001**, *52*, 499–535.

(16) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 826–843.

(17) Dickson, A.; Brooks III, C. L. WExplore: hierarchical exploration of high-dimensional spaces using the weighted ensemble algorithm. *J. Phys. Chem. B* **2014**, *118*, 3532–3542.

(18) Zwier, M. C.; Adelman, J. L.; Kaus, J. W.; Pratt, A. J.; Wong, K. F.; Rego, N. B.; SuaÌĄrez, E.; Lettieri, S.; Wang, D. W.; Grabe, M.; Zuckerman, D. M.; Chongothers, L. T. WESTPA: An interoperable, highly scalable software package for weighted ensemble simulation and analysis. *J. Chem. Theory Comput.* **2015**, *11*, 800–809.

(19) Machta, J.; Ellis, R. S. Monte Carlo methods for rough free energy landscapes: Population annealing and parallel tempering. *J. Stat. Phys.* **2011**, *144*, 541–553.

(20) Zhang, J.; Li, W.; Wang, J.; Qin, M.; Wang, W. All-atom replica exchange molecular simulation of protein BBL. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 1038–1047.

(21) Kannan, S.; Zacharias, M. Folding simulations of Trp-cage mini protein in explicit solvent using biasing potential replica-exchange molecular dynamics simulations. *Proteins: Struct., Funct., Bioinf.* **2009**, *76*, 448–460.

(22) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116*, 9058–9067.

(23) Affentranger, R.; Tavernelli, I.; Di Iorio, E. E. A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling. *J. Chem. Theory Comput.* **2006**, *2*, 217–228.

(24) Kannan, S.; Zacharias, M. Enhanced sampling of peptide and protein conformations using replica exchange simulations with a peptide backbone biasing-potential. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 697–706.

(25) Sabri Dashti, D.; Roitberg, A. E. Optimization of Umbrella Sampling Replica Exchange Molecular Dynamics by Replica Positioning. *J. Chem. Theory Comput.* **2013**, *9*, 4692–4699.

(26) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **2010**, *52*, 99–105.

(27) Shakhnovich, E.; Farztdinov, G.; Gutin, A.; Karplus, M. Protein folding bottlenecks: A lattice Monte Carlo simulation. *Phys. Rev. Lett.* **1991**, *67*, 1665.

(28) Shea, J.-E.; Onuchic, J. N.; Brooks III, C. L. Energetic frustration and the nature of the transition state in protein folding. *J. Chem. Phys.* **2000**, *113*, 7663–7671.

(29) Hardin, C.; Luthey-Schulten, Z.; Wolynes, P. G. Backbone dynamics, fast folding, and secondary structure formation in helical proteins and peptides. *Proteins: Struct., Funct., Bioinf.* **1999**, *34*, 281–294.

(30) Eastwood, M. P.; Wolynes, P. G. Role of explicitly cooperative interactions in protein folding funnels: a simulation study. *J. Chem. Phys.* **2001**, *114*, 4702–4716.

(31) Pogorelov, T. V.; Luthey-Schulten, Z. Variations in the fast folding rates of the λ-repressor: A hybrid molecular dynamics study. *Biophys. J.* **2004**, *87*, 207–214.

(32) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E.; Mittal, J.; Feig, M.; MacKerell Jr, A. D. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi1$ and $\chi2$ dihedral angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257–3273.

(33) Sheinerman, F. B.; Brooks, C. L. Molecular picture of folding of a small α/β protein. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 1562–1567.

(34) Chen, J.; Brooks, C. L. Can molecular dynamics simulations provide high-resolution refinement of protein structure? *Proteins: Struct., Funct., Bioinf.* **2007**, *67*, 922–930.

(35) Best, R. B.; Paci, E.; Hummer, G.; Dudko, O. K. Pulling direction as a reaction coordinate for the mechanical unfolding of single molecules. *J. Phys. Chem. B* **2008**, *112*, 5968–5976.

(36) Levy, Y.; Becker, O. M. Energy landscapes of conformationally constrained peptides. *J. Chem. Phys.* **2001**, *114*, 993–1009.

(37) Vengadesan, K.; Gautham, N. Energy landscape of Met-enkephalin and Leu-enkephalin

drawn using mutually orthogonal Latin squares sampling. *J. Phys. Chem. B* **2004**, *108*, 11196–11205.

(38) Itoh, K.; Sasai, M. Flexibly varying folding mechanism of a nearly symmetrical protein: B domain of protein A. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 7298–7303.

(39) Guo, W.; Lampoudi, S.; Shea, J.-E. Posttransition state desolvation of the hydrophobic core of the src-SH3 protein domain. *Biophys. J.* **2003**, *85*, 61–69.

(40) Juraszek, J.; Bolhuis, P. G. Rate constant and reaction coordinate of Trp-cage folding in explicit water. *Biophys. J.* **2008**, *95*, 4246–4257.

(41) Jiang, F.; Wu, Y.-D. Folding of fourteen small proteins with a residue-specific force field and replica-exchange molecular dynamics. *J. Am. Chem. Soc.* **2014**, *136*, 9536–9539.

(42) Best, R. B.; Hummer, G.; Eaton, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 17874–17879.

(43) Guo, W.; Lampoudi, S.; Shea, J.-E. Temperature dependence of the free energy landscape of the src-SH3 protein domain. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 395–406.

(44) Bursulaya, B. D.; Brooks, C. L. Folding free energy surface of a three-stranded β-sheet protein. *J. Am. Chem. Soc.* **1999**, *121*, 9947–9951.

(45) Sun, L.; Noel, J. K.; Sulkowska, J. I.; Levine, H.; Onuchic, J. N. Connecting thermal and mechanical protein (un) folding landscapes. *Biophys. J.* **2014**, *107*, 2950–2961.

(46) Meuzelaar, H.; Marino, K. A.; Huerta-Viga, A.; Panman, M. R.; Smeenk, L. E.; Kettelarij, A. J.; van Maarseveen, J. H.; Timmerman, P.; Bolhuis, P. G.; Woutersen, S. Folding dynamics of the Trp-cage miniprotein: evidence for a native-like intermediate

from combined time-resolved vibrational spectroscopy and molecular dynamics simulations. *J. Phys. Chem. B* **2013**, *117*, 11490–11501.

(47) Sarisky, C. A.; Mayo, S. L. The $\beta\beta\alpha$ fold: explorations in sequence space. *J. Mol. Biol.* **2001**, *307*, 1411–1418.

(48) Juraszek, J.; Bolhuis, P. Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 15859–15864.

(49) Marinelli, F.; Pietrucci, F.; Laio, A.; Piana, S. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. *PLoS Comput. Biol.* **2009**, *5*, e1000452.

(50) Barua, B.; Lin, J. C.; Williams, V. D.; Kummler, P.; Neidigh, J. W.; Andersen, N. H. The Trp-cage: optimizing the stability of a globular miniprotein. *Protein Eng., Des. Sel.* **2008**, *21*, 171–185.

(51) MacKerell Jr, A. D.; Feig, M.; Brooks, C. L. Improved treatment of the protein backbone in empirical force fields. *J. Am. Chem. Soc.* **2003**, *126*, 698–699.

(52) MacKerell Jr, A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K. K.; Lau, F. T. K.; Mattos, C.; Michnick, S. K.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R. H.; Straub, J. E.; Watanabe, M.; WiÃşrkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteinsâĂă. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(53) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(54) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(55) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant pressure molecular dynamics simulation: the Langevin piston method. *J. Chem. Phys.* **1995**, *103*, 4613–4621.

(56) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

(57) Miyamoto, S.; Kollman, P. A. SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952–962.

(58) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An NâŃĚ log (N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(59) Park, S.; Schulten, K. Calculating potentials of mean force from steered molecular dynamics simulations. *J. Chem. Phys.* **2004**, *120*, 5946–5961.

(60) Zhu, F.; Hummer, G. Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *J. Comput. Chem.* **2012**, *33*, 453–465.

(61) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(62) Neale, C.; Madill, C.; Rauscher, S.; Pomès, R. Accelerating convergence in molecular dynamics simulations of solutes in lipid membranes by conducting a random walk along the bilayer normal. *J. Chem. Theory Comput.* **2013**, *9*, 3686–3703.

(63) Zhu, F.; Hummer, G. Theory and simulation of ion conduction in the pentameric GLIC channel. *J. Chem. Theory Comput.* **2012**, *8*, 3759–3768.

(64) Song, H. D.; Zhu, F. Finite Temperature String Method with Umbrella Sampling: Application on a Side Chain Flipping in Mhp1 Transporter. *J. Phys. Chem. B* In press.

(65) Hummer, G. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.* **2005**, *7*, 34.

(66) Bolen, D. W.; Rose, G. D. Structure and energetics of the hydrogen-bonded backbone in protein folding. *Annu. Rev. Biochem.* **2008**, *77*, 339–362.

(67) Zhou, C.-Y.; Jiang, F.; Wu, Y.-D. Folding thermodynamics and mechanism of five TRP-cage variants from replica-exchange MD simulations with RSFF2 force field. *J. Chem. Theory Comput.* **2015**, *11*, 5473–5480.

(68) Hałabis, A.; ZÌĞmudzinÌĄska, W.; Liwo, A.; Ołdziej, S. Conformational dynamics of the trp-cage miniprotein at its folding temperature. *J. Phys. Chem. B* **2012**, *116*, 6898–6907.

(69) Ahmed, Z.; Beta, I. A.; Mikhonin, A. V.; Asher, S. A. UV-resonance Raman thermal unfolding study of Trp-cage shows that it is not a simple two-state miniprotein. *J. Am. Chem. Soc.* **2005**, *127*, 10943–10950.

(70) Culik, R. M.; Serrano, A. L.; Bunagan, M. R.; Gai, F. Achieving Secondary Structural Resolution in Kinetic Measurements of Protein Folding: A Case Study of the Folding Mechanism of Trp-cage. *Angew. Chem., Int. Ed.* **2011**, *50*, 10884–10887.

(71) Marinelli, F. Following easy slope paths on a free energy landscape: The case study of the Trp-cage folding mechanism. *Biophys. J.* **2013**, *105*, 1236–1247.

(72) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. Designing a 20-residue protein. *Nat. Struct. Mol. Biol.* **2002**, *9*, 425–430.

(73) Paschek, D.; Hempel, S.; García, A. E. Computing the stability diagram of the Trp-cage miniprotein. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 17754–17759.

(74) Day, R.; Paschek, D.; Garcia, A. E. Microsecond simulations of the folding/unfolding thermodynamics of the Trp-cage miniprotein. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1889–1899.

(75) Piana, S.; Klepeis, J. L.; Shaw, D. E. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98–105.

(76) Nettels, D.; Müller-Späth, S.; Küster, F.; Hofmann, H.; Haenni, D.; Rüegger, S.; Reymond, L.; Hoffmann, A.; Kubelka, J.; Heinz, B.; Gast, K.; Best, R. B.; Schulera, B. Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 20740–20745.

(77) Henriques, J.; Cragnell, C.; SkepoÌĹ, M. Molecular dynamics simulations of intrinsically disordered proteins: Force field evaluation and comparison with experiment. *J. Chem. Theory Comput.* **2015**, *11*, 3420–3431.

(78) Best, R. B.; Zheng, W.; Mittal, J. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **2014**, *10*, 5113–5124.

(79) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **2011**, *100*, L47–L49.

(80) Debiec, K. T.; Gronenborn, A. M.; Chong, L. T. Evaluating the strength of salt bridges:

a comparison of current biomolecular force fields. *J. Phys. Chem. B* **2014**, *118*, 6561–6569.

(81) Pietrucci, F.; Mollica, L.; Blackledge, M. Mapping the native conformational ensemble of proteins from a combination of simulations and experiments: new insight into the src-SH3 domain. *J. Phys. Chem. Lett.* **2013**, *4*, 1943–1948.

Figures

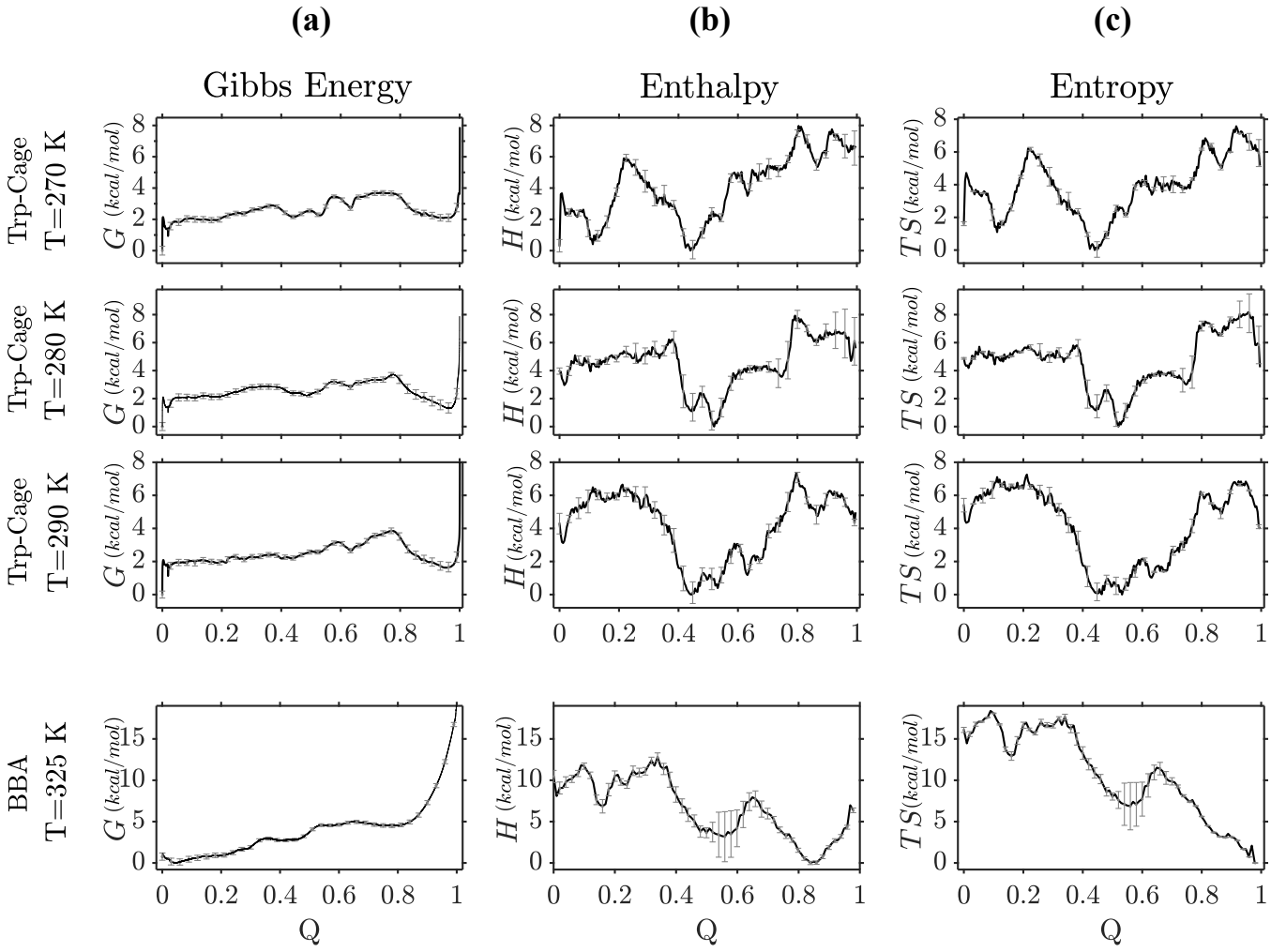

Figure 1: Energetics along the reaction coordinate Q from the US simulations. a) The free energy profiles calculated from the WHAM[60,61] equations. The statistical errors are with respect to the difference between the free energy value at the given position and the average value of the entire profile, and were estimated from the uncertainties in the mean force at each umbrella window.[60] b) The profile of average enthalpy along Q. c) The entropy multiplied by the temperature.

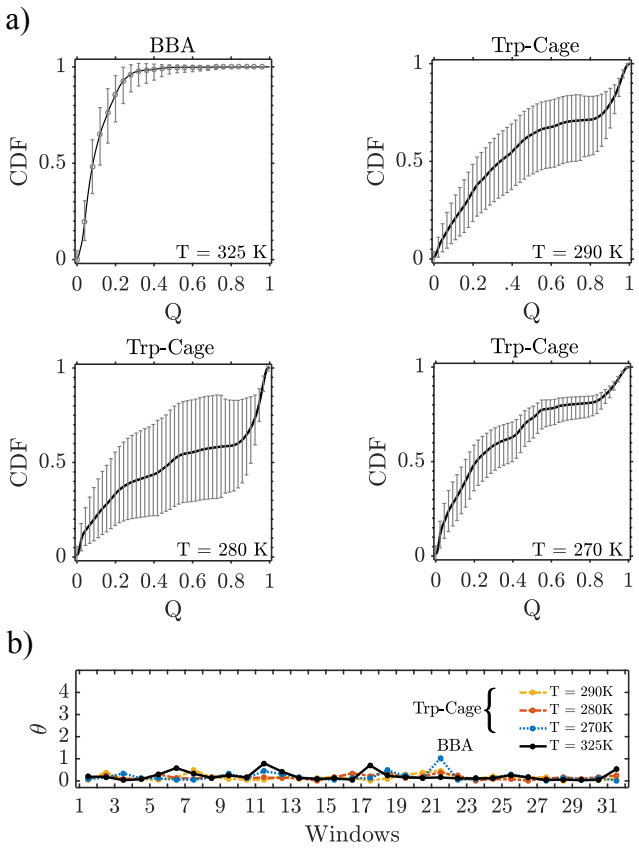

Figure 2: a) Cumulative distribution function obtained by integrating the equilibrium probability distribution along Q. The error bars at each data point were estimated separately. For any given point Q_i, the upper and lower bounds (taken as ± 1 standard deviation) for the profile of the free energy differences relative to Q_i were obtained (similarly from the statistical errors in the mean force for each window) and used to calculate the upper and lower limits for the cumulative probability at Q_i. b) Inconsistency coefficient θ for pairs of histograms in the adjacent umbrella windows.[60]

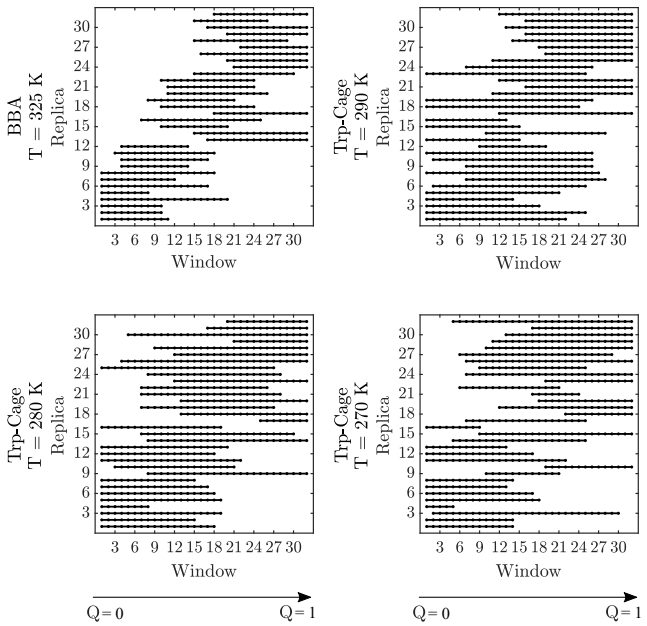

Figure 3: The umbrella windows that each replica sampled during the second half of the US simulations.

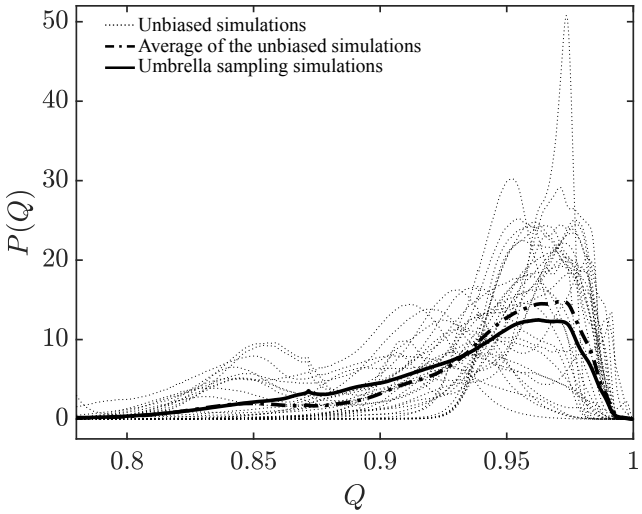

Figure 4: Data from the 32 unbiased simulations (344 ns each) at the native state of Trp-Cage at 280 K. For each unbiased simulation, the histogram from the second half (172 ns) of the trajectory is shown as a *dotted* line. The average of the 32 histograms is shown as the *dashed* line. The *solid* line shows the normalized equilibrium probabilities for the range of Q representing the native conformation, which were calculated from the US simulations (cf. figure 1a).

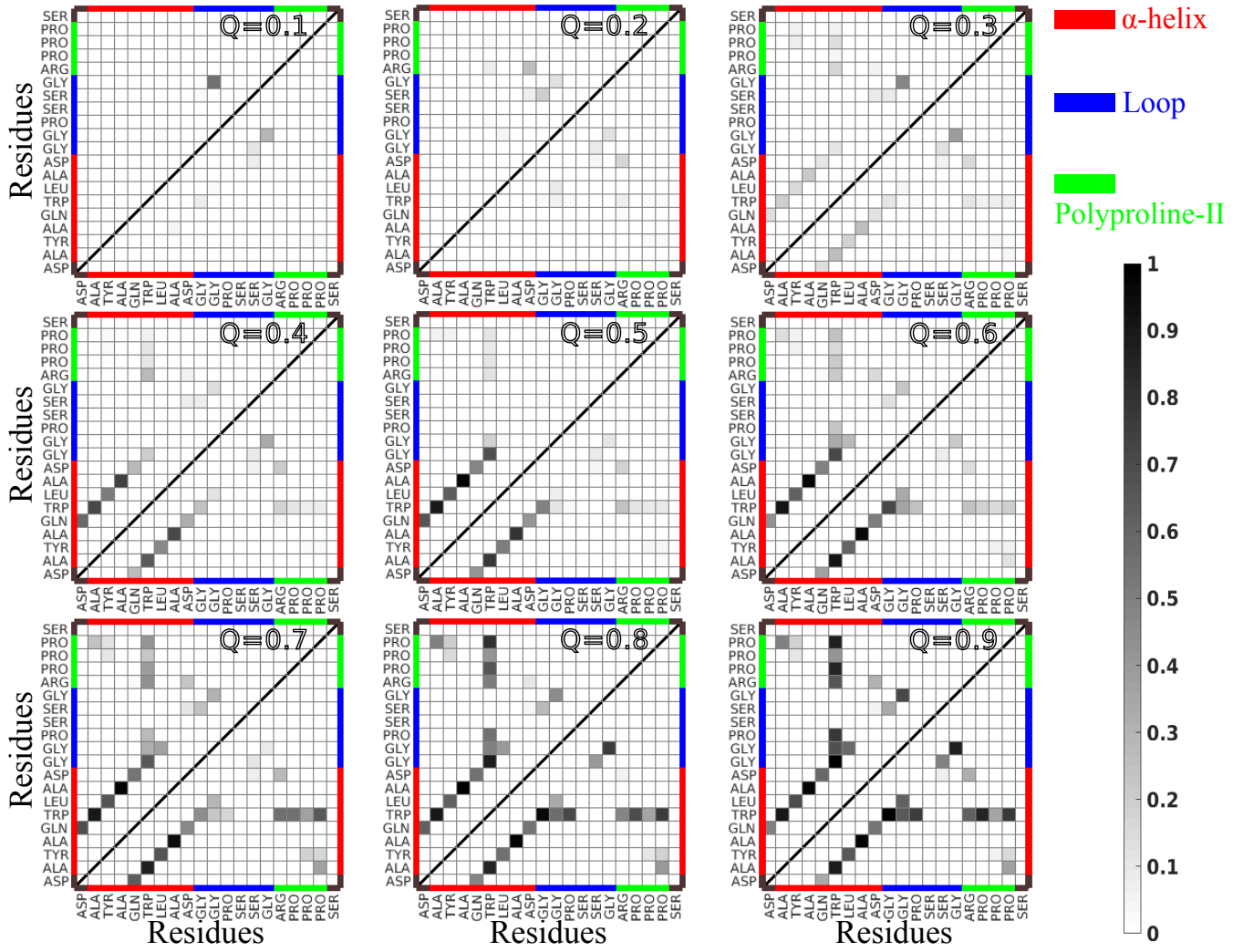

Figure 5: The fraction of the native contacts (or the average contact strength) between each pair of residues in the Trp-Cage conformations with different Q values at 270 K (*upper left*) and 290 K (*lower right*). For each Q value, conformations within $Q \pm 0.01$ were taken to calculate the average contact strength between every residue pair in the protein.

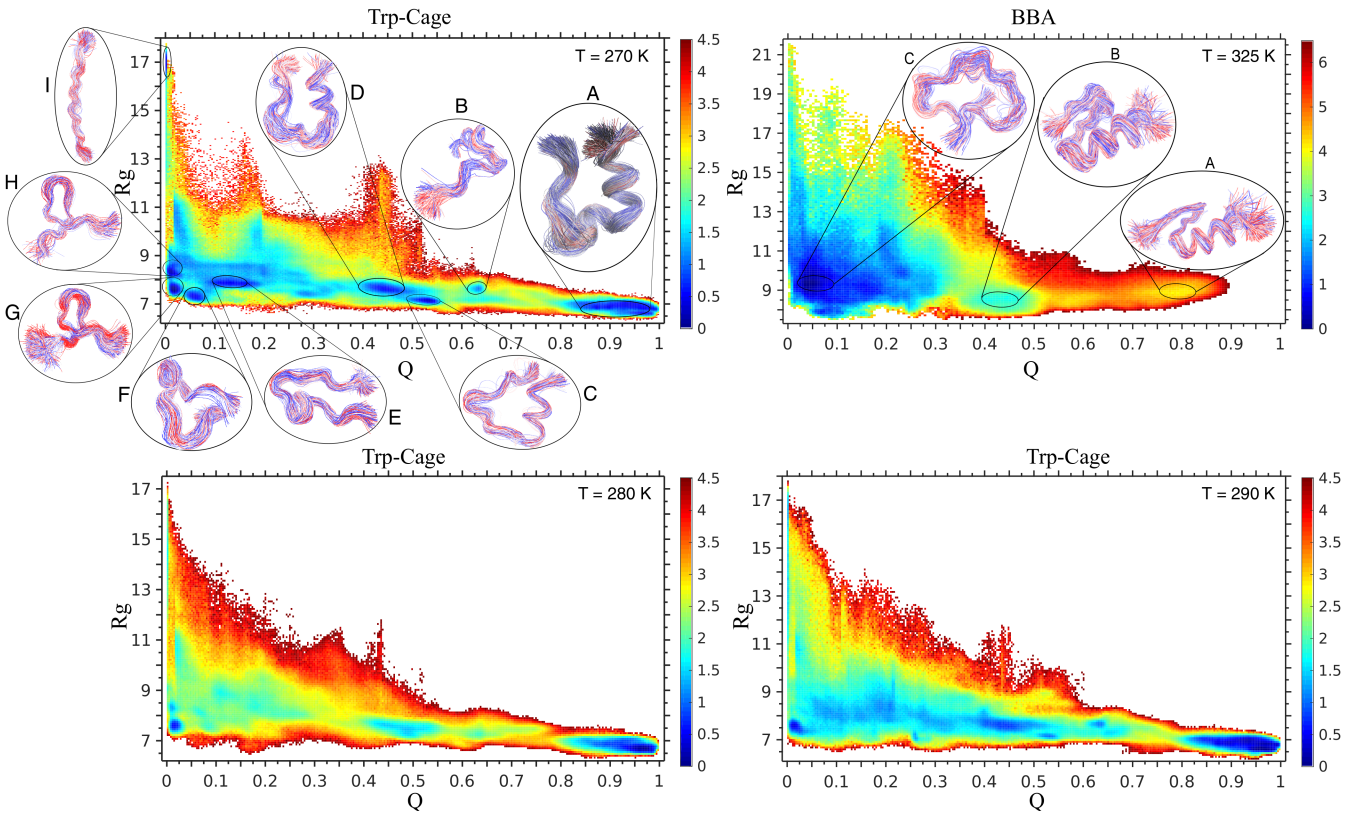

Figure 6: Two-dimensional free energy (in unit of kcal/mol) maps as a function of the reaction coordinate (Q) and the radius of gyration (Rg, in unit of Å) of the protein conformation, for Trp-Cage at 270 K, 280 K, and 290 K and BBA at 325 K. The free energies were determined from the joint probability distribution of Q and Rg in the equilibrium ensemble. Some representative conformations at various free energy minima are also shown in the figure.

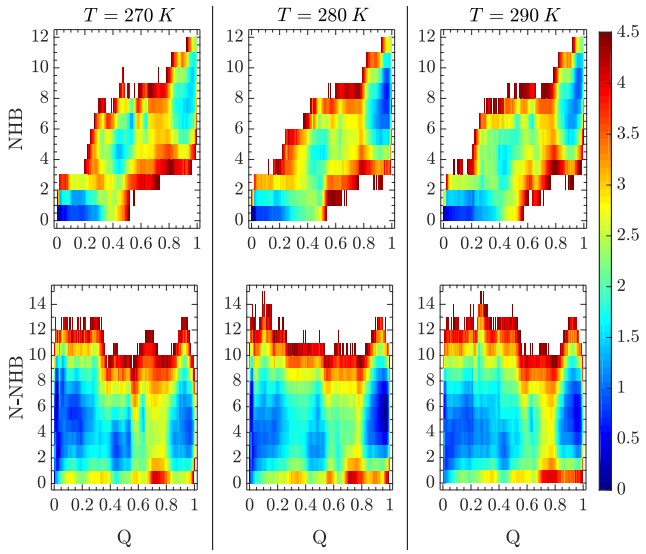

Figure 7: Two-dimensional free energy (in unit of kcal/mol) maps as a function of Q and the number of NHBs (*first row*) or the number of N-NHBs (*second row*) for Trp-Cage at 270 K (*left*), 280 K (*middle*) and 290 K (*right*). The free energies were determined from the joint probability distribution of Q and the H-bond count in the equilibrium ensemble.

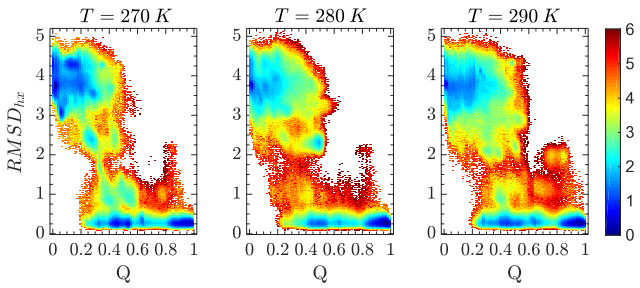

Figure 8: Two-dimensional free energy (in unit of kcal/mol) maps as a function of Q and the C_α RMSD (in unit of Å) for the α-helix (residue 2–9) in Trp-Cage at 270 K (*left*), 280 K (*middle*) and 290 K (*right*). The free energies were determined from the joint probability distribution of Q and the RMSD in the equilibrium ensemble.

Graphical TOC Entry